Analysis report of Auckland house prices

Gia Han To

24/07/2020

Executive Summary

The dataset contains information about property in Auckland city from different suburbs. Information includes number of bedrooms, bathrooms in each individual property, physical address of the property, land area of property in meters squared, capital value of the property, latitude and longitude coordinate of the property, SA1- area unit classification, number of people living in the property range from age 0 to 60+, and lastly suburb of the property. Two attributes are later added to the dataset which are the population of each suburb and the deprivation index .

The response variable is CV – capital value of the property which is an approximation of house value. There are two categorical variables in the dataset: Address and Suburbs. Address is dropped as they are strings, and there should not be any pattern in these strings can affect the house price. Latitude, Longitude, SA1 are also dropped as they are not relevant to the model we are building.

The analysis is based on 1050 observations for each of the 13 variables. The rest of the variables is explanatory variables. They have an impact on the house price in Auckland. All the measurements have mean, standard error and worst and variables.

After exploring the data by calculating summary and descriptive statistics, and by creating visualizations of the correlation between each numerical variable, it shows no strong correlation relationship. After exploring the data, three algorithms have been tested for the train dataset and the best model has been chosen based on the model accuracy.

Initial Data Exploration

The initial exploration of the data began with some summary and descriptive statistics.

Individual Feature Statistics Summary statistics for minimum, maximum, mean, median, standard deviation, and distinct count were calculated for numeric columns.

```
##     Bedrooms       Bathrooms
##  Min.   : 1.00   Min.   :0.00
##  1st Qu.: 3.00   1st Qu.:1.00
##  Median : 4.00   Median :2.00
```

```
##  Mean   : 3.78   Mean   :2.07
##  3rd Qu.: 4.00   3rd Qu.:3.00
##  Max.   :17.00   Max.   :8.00
##                                                  Address       Land.area
##  11 Marie Costello Way Beach Haven, Auckland :   2   Min.   :   40.0
##  118 Reihana Street Orakei, Auckland         :   2   1st Qu.:  320.0
##  16 Percy Winstone Lane Stonefields, Auckland:   2   Median :  570.5
##  46 Waiatarua Road Remuera, Auckland         :   2   Mean   :  855.8
##  65 Old Coach Way Ramarama, Auckland         :   2   3rd Qu.:  825.0
##  8 Justamere Place Weymouth, Auckland        :   2   Max.   :22240.0
##  (Other)                                     :1038
##        CV              Latitude        Longitude          SA1
##  Min.   :  270000   Min.   :-37.27   Min.   :174.3   Min.   :7001130
##  1st Qu.:  780000   1st Qu.:-36.95   1st Qu.:174.7   1st Qu.:7004418
##  Median : 1080000   Median :-36.89   Median :174.8   Median :7006329
##  Mean   : 1388137   Mean   :-36.89   Mean   :174.8   Mean   :7006324
##  3rd Qu.: 1600000   3rd Qu.:-36.86   3rd Qu.:174.9   3rd Qu.:7008384
##  Max.   :18000000   Max.   :-36.18   Max.   :175.5   Max.   :7011028
##
##   X0.19.years      X20.29.years     X30.39.years     X40.49.years
##  Min.   :  0.00   Min.   :  0.00   Min.   :  0.00   Min.   :  0.00
##  1st Qu.: 33.00   1st Qu.: 15.00   1st Qu.: 15.00   1st Qu.: 18.00
##  Median : 45.00   Median : 24.00   Median : 24.00   Median : 24.00
##  Mean   : 47.57   Mean   : 28.99   Mean   : 27.06   Mean   : 24.13
##  3rd Qu.: 57.00   3rd Qu.: 36.00   3rd Qu.: 33.00   3rd Qu.: 30.00
##  Max.   :201.00   Max.   :270.00   Max.   :177.00   Max.   :114.00
##
##   X50.59.years    X60..years              Suburbs      Population
##  Min.   : 0.0   Min.   :  0.00   Remuera      : 61   Min.   :  3.0
##  1st Qu.:15.0   1st Qu.: 18.00   Manurewa     : 38   1st Qu.:138.0
##  Median :21.0   Median : 27.00   Papatoetoe   : 29   Median :174.0
##  Mean   :22.6   Mean   : 29.33   St Heliers   : 29   Mean   :179.9
##  3rd Qu.:27.0   3rd Qu.: 36.00   Mount Eden   : 26   3rd Qu.:210.0
##  Max.   :90.0   Max.   :483.00   Mount Roskill: 26   Max.   :789.0
##                                  (Other)      :841
##    NZDep2018
##  Min.   : 1.00
##  1st Qu.: 2.00
##  Median : 5.00
##  Mean   : 5.06
##  3rd Qu.: 8.00
##  Max.   :10.00
##

##     Min.  1st Qu.   Median     Mean  3rd Qu.     Max.
##   270000   780000  1080000  1388137  1600000 18000000
```

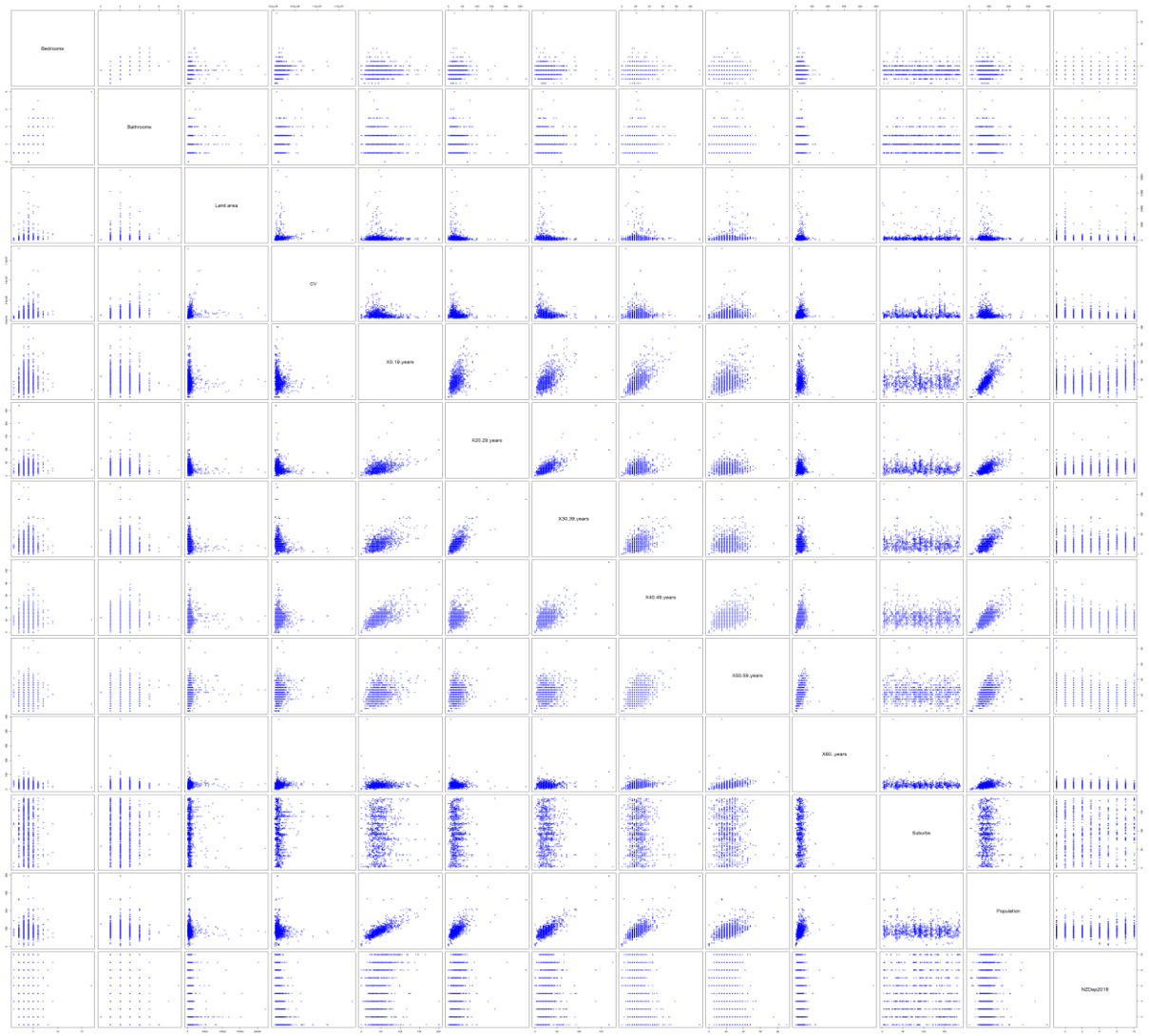
#Median is markedly lower than the mean value.CV shows a right-skewed distribution which is a very common when it comes to thing involving money, growth, salary and age.

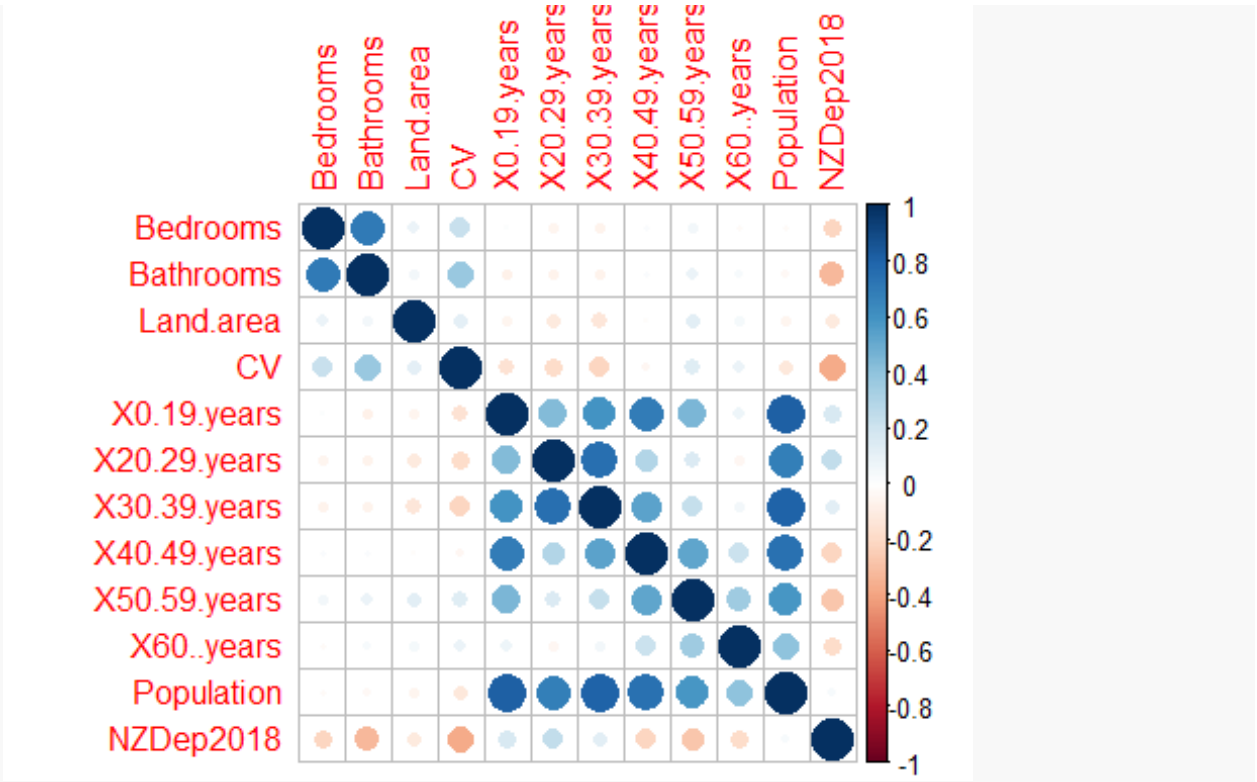
Analysis of correlations and patterns

The pair plot below shows that CV and deprivation index has a decreasing linear relationship. The smaller the deprivation index is, the higher the house price will be.

The correlation between numeric columns were calculated and observed in the below correlation plot. (The right color bar indicated the correlation values. For example, dark blue correlation value is 1 and dark red means correlation value is negative 1)

The correlation plot shows no strong correlated relationship between numeric variables. CV and Bathrooms has a slightly more positively correlated relationship compared to others. CV and deprivation index is negatively correlated with each other.





Analysis

In this analysis, three algorithms have been tested, which are linear regression model, linear regression model with log-transformed data and linear regression model with log-transformed data and dropping irrelevant attributes.

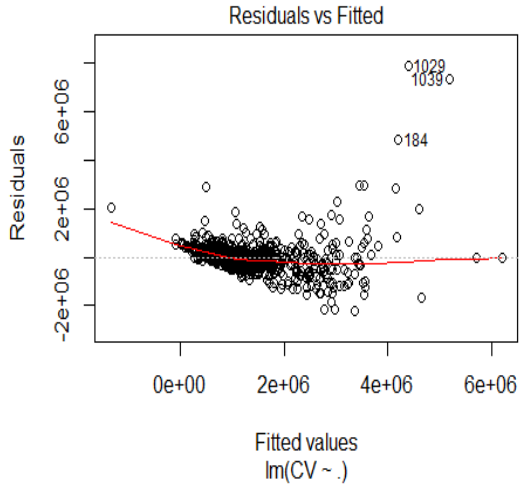
These algorithms are trained with 70% of the data. Testing the model with the remaining 30% of the data.

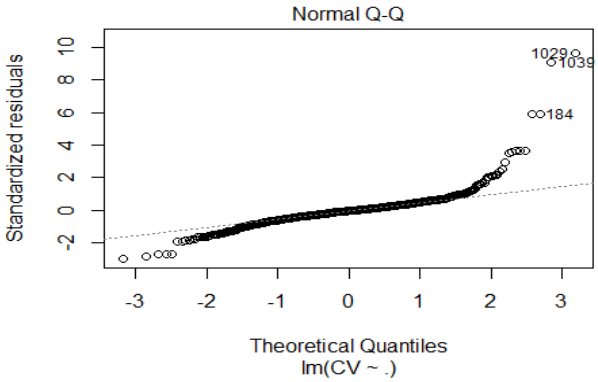
Model 1 – Linear Regression Model

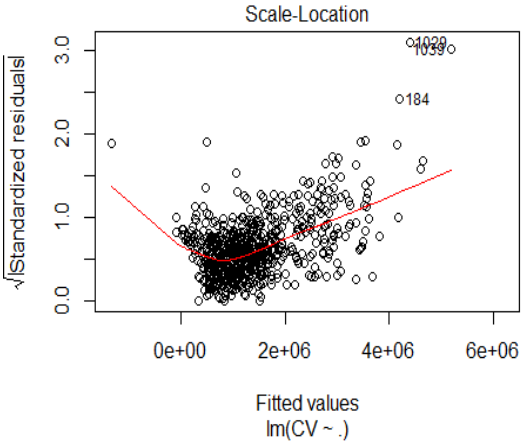
The residual vs fitted plot of model 1 shows clusters and a slight pattern. This can be due to the right-skewed distribution of the CV. According to the summary of CV, its median is rem arkedly lower than the mean value. This is very common when it comes to things involving money, prices, growth, salary and age. The QQ plot shows residuals are normally distribute d, there are few potential outliers shown but it looks quite good overall.

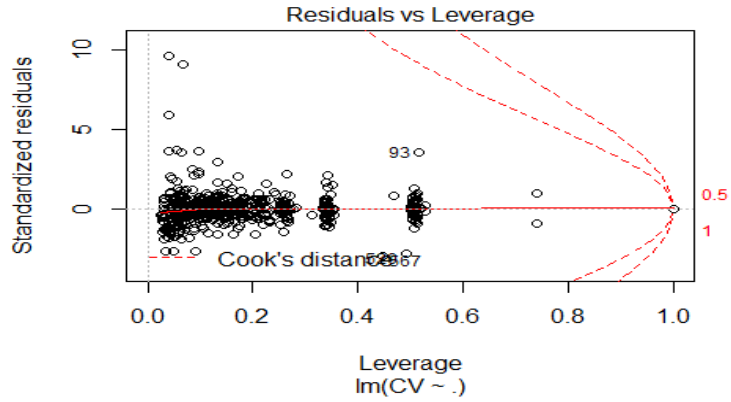
There seems to be a linear relationship between CV and other explanatory variables (P-val ue ≈ 0).

The model explains 59% of variability in the capital value of the property.


Residuals vs Fitted
Residuals
Fitted values
lm(CV ~ .)
1029
1039
184
-2e+06
2e+06
6e+06
0e+00
2e+06
4e+06
6e+06


Normal Q-Q
Standardized residuals
Theoretical Quantiles
lm(CV ~ .)
1029
1039
184
-2
0
2
4
6
8
10
-3
-2
-1
0
1
2
3


Scale-Location
√|Standardized residuals|
Fitted values
lm(CV ~ .)
1029
1039
184
0.0
1.0
2.0
3.0
0e+00
2e+06
4e+06
6e+06


Residuals vs Leverage
Standardized residuals
Leverage
lm(CV ~ .)
Cook's distance
93
0.5
1
0
5
10
0.0
0.2
0.4
0.6
0.8
1.0

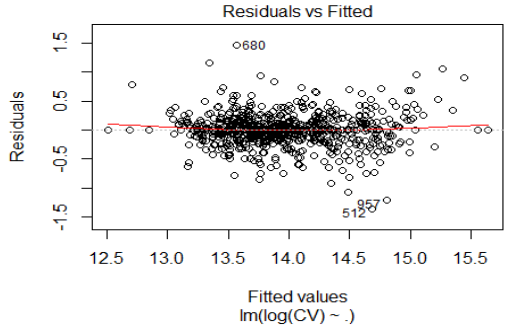
Model 2 – Linear Regression Model with log-transformed data

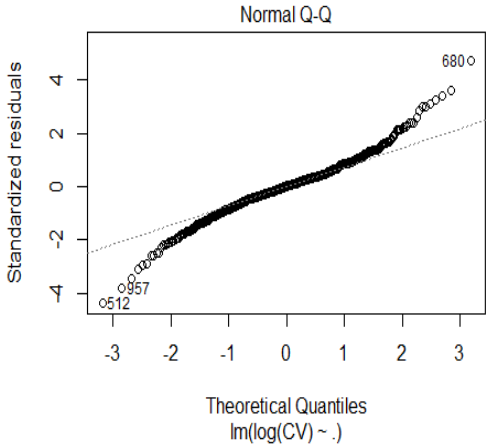
For model 2, CV is log transformed to account for the pattern shown in residuals. After logging the data, residuals are more spread out along the horizontal line without distinct pattern. The QQ plot indicates that residuals are normally distributed, however, there are more outliers spotted in model 2 than model 1. This should be taken into the consideration of whether model 2 is indeed a better fit than model 1.

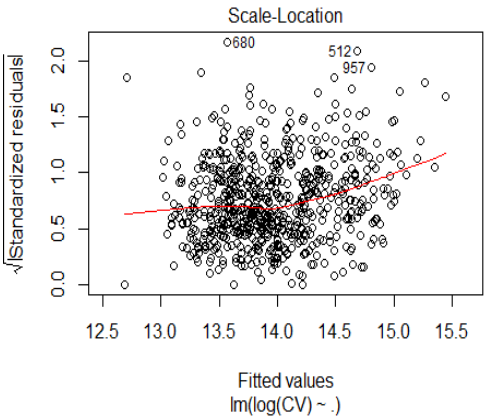
There seems to be a linear relationship between CV and other explanatory variables (P-val ue ≈ 0).

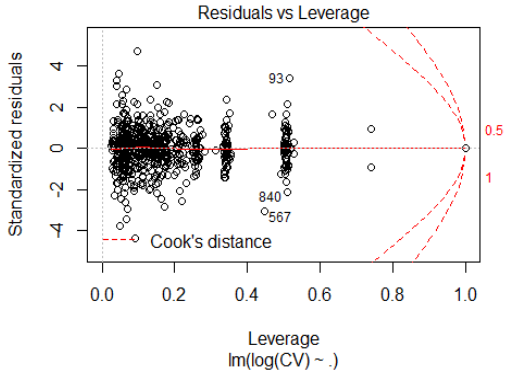
The model explains 74% of variability in the logged capital value of the property.

```
## Residual standard error: 0.3276 on 556 degrees of freedom
## Multiple R-squared:  0.744,  Adjusted R-squared:  0.662
## F-statistic: 9.078 on 178 and 556 DF,  p-value: < 2.2e-16
```









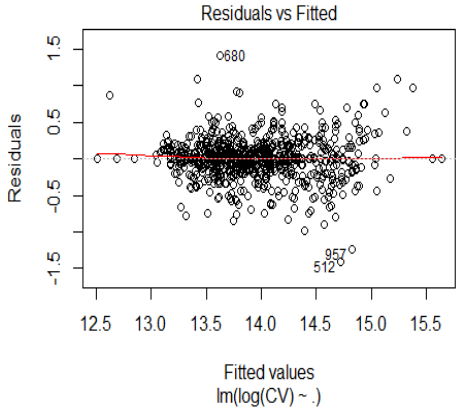
Model 3 – Keeping key variables

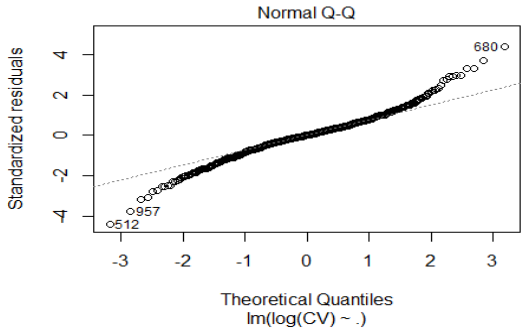
In model 3, non-significant values are dropped. There is not much of an improvement from model 2 to model 3. All the plots look similar to model 2.

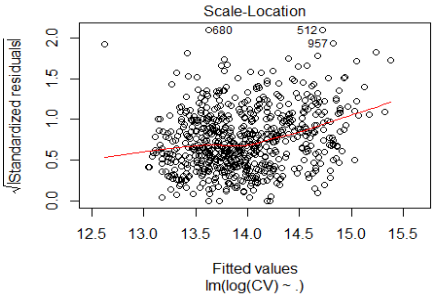
There seems to be a linear relationship between CV and other explanatory variables (P-val ue ≈ 0).

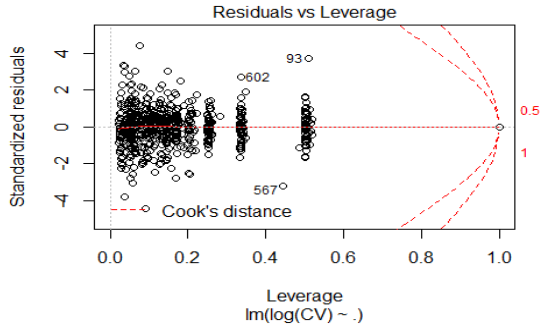
Model 3 explains 73% variability in the logged capital value of property.

```
## Residual standard error: 0.3336 on 562 degrees of freedom
## Multiple R-squared:  0.7316, Adjusted R-squared:  0.6494
## F-statistic: 8.904 on 172 and 562 DF,  p-value: < 2.2e-16
```










```
#Prediction
test = test[test$Suburbs %in% train$Suburbs, , drop=FALSE]

prediction1 = predict(model1,test)
prediction2 = predict(model2,test)
prediction3 = predict(model3,test)

#Accuracy
actual = test$CV
RMSE1 = sqrt(mean(prediction1-actual)^2)

## [1] 77513.11

RMSE2 = sqrt(mean(exp(prediction2)-actual)^2)

## [1] 125104.9

RMSE3 = sqrt(mean(exp(prediction3)-actual)^2)

## [1] 131829.7
```

Conclusion

Model 1 gives the smallest RMSE out of all three algorithms. This brings to the conclusion that model 1 is the best model to use on the dataset. However, model 1 only explains 56%

of variability in the capital value. More algorithms should be considered and try to better fit the dataset.